Cross-situational inference and meaning space structure

B005318
MSc Evolution of Language & Cognition
The University of Edinburgh
2011

Contents

> In the traditional analyses of ‘‘concept’’ and ‘‘reference’’ that dominate thinking in cognitive science, ‘‘concept’’ is treated as a noun akin to dog, and ‘‘refer’’ as a verb akin to point…An alternative characterization [is] that concepts are better understood in terms of the probabilistic knowledge used in making predictions (rather than as discrete units of representation), and that reference should be understood in terms of the way that words function as tools in a predictive human practice (rather than as pointers to the world). These proposals are not new ways of approaching traditional problems in cognitive science, but rather represent a fundamental reanalysis of the problems cognitive science seeks to solve. (Ramscar et al. 2010)

1. Introduction

Humans are the only species who use a vocal communication system to flexibly direct each other's attention to chosen aspects of the world. By combining words systematically, we can bring our hearer to understand a complex meaning: 'simply by making noises with our mouths, we can reliably cause precise new combinations of ideas to arise in each other's minds' (Pinker 1994 p.15).

The meanings that can be communicated in this way are vast, limited only by the speaker's experience and lexical knowledge. If I want you to visualise Niagara Falls, I don't have to take you there, I can use words to describe how it looks. If I want you to understand that I would like you to pass me the salt, I don't have to point at it and look mournful, I can simply ask you. If I want you to know that yesterday the cooker exploded while you were out of the house, I don't have to invent time travel so you can be present at the event, I can say, 'The cooker exploded'. This capacity to communicate wide-ranging, flexible meanings that can encompass external states of affairs, internal attitudes and non-present events is far beyond the ability of any other species, which are mainly limited to situation-specific signalling repertoires which 'lack the rich expressive and open-ended power of human language' (Hauser, Chomsky & Fitch 2002).

Researchers who study the origins of language are faced with the problem of explaining how a human ancestral species could have passed from a language-less state, perhaps no more complex than the call repertoires of extant non-human primates, to human language, with its ability to express a wide range of meanings in a wide range of ways. Where did these meanings come from? Were they pre-established in individual minds before the advent of language, just awaiting the arrival

of words to match them? Words, as mentioned above, come in recombinable units: does this imply that the meanings of words are cognitively structured in the same way, and if so, does this structure come from the world, the mind, or from the process of communication itself?

Researchers have begun to investigate these questions by hypothesising a transitional phase or 'protolanguage' that preceded full language as we know it, in which only a limited number of meanings were expressible. When considering how these meanings were first grounded, an inextricably linked question is what kind of meanings they were likely to be. Here, protolanguage researchers have split into two main schools of thought. The synthetic model (Bickerton 1990, Tallerman 2007) assumes a 'word-based' protolanguage, whose meanings corresponded to 'pre-existing cognitive concepts – the prototypes of nouns and verbs' (Tallerman 2007 p.580). The assumption is that pre-linguistic cognitive concepts naturally corresponded to the referents of present-day nouns and verbs. The holistic model (Wray 2002) argues that the meanings of the first utterances were less like modern-day nouns and verbs and instead were complex, socially manipulative messages intended to influence the behaviour of conspecifics. Wray (2000) argues that this approach allows a closer link to the call systems of non-human primates. Her theory then suggests a process whereby these initially complex meanings were later broken down into atomic meanings that could be flexibly recombined.

Before we decide which of these theories provides the most plausible account of the evolution of language, it pays first to take a closer look at the meanings associated with language-like behaviour in extant non-human primates. How do we establish what a primate call means? The obvious answer is to look at what precedes and follows the call, to try and infer from this evidence what triggers the calling behaviour and how the listeners respond. This is what Seyfarth et al. (1980) do in their classic study of alarm calls in vervet monkeys.

In the presence of a leopard, vervets tend to make a particular short tonal call. When a vervet monkey makes this call, the other vervets in the group run up a tree, a well-adapted response to this specific danger (Seyfarth et al. 1980, p.802). Seyfarth et al. talk about the vervets' use of alarm calls in terms of reference, or 'the systematic use of signals to refer to objects in the external world' (p.801). However, their study does not show that the call functions referentially for the vervets. All its shows is that the alarm call is reliably associated with a specific response. This leaves the semantics

of the call unclear: it could just as well be an imperative with the more semantically complex meaning 'Run up a tree now!', or a warning incorporating both reference to the leopard and an injunction to beware.

Further examination of the vervet example may illuminate why both of the approaches sketched above may be going at the problem the wrong way. Jim Hurford, working from the synthetic side of the argument, claims that 'the English speaker's concept associated with the word *leopard* has a lot in common with what the vervet has in its head that makes it respond systematically to sight of a leopard or sound of a warning bark', although the vervet concept 'also triggers (or includes?) the typical motor response of running up a tree' (Hurford 2007, p.97). However, an English-speaker can use the word 'leopard' in a variety of contexts: not only to alert others to the presence of a leopard, but also to talk about a leopard they once saw at a zoo or to compare a pattern to a leopard's skin. It is what remains constant across all these contexts (i.e. various aspects of memory or sensory input associated with the animal we call a leopard) that both continually re-grounds the meaning, and allows the word to continue being communicatively useful.

Conversely, for the vervets, because of the lack of variation in context of use, it is impossible to pin down the concept associated with the call as LEOPARD, BEWARE, RUN-UP-TREE, or a combination of the above. The reason is that the calls are never used outside a context where there is both a leopard present and a need to run up a tree. This makes evolutionary sense, since vervets who sat around discussing the presence of a leopard without responding to the danger would not survive to reproduce. However, it also makes comparison to the referential, context-sensitive nature of human communication problematic. Monkey calls, unlike human words, cannot be detached from one particular context and re-used flexibly. In the words of Ray Jackendoff, 'a food call is used when food is discovered (or imminently anticipated), but not to suggest that food be sought. A leopard alarm call can report the sighting of a leopard, but cannot ask if anyone has seen a leopard lately' (Jackendoff 1999 p.273). The logical question that follows is: what is unique about human communication that allows flexible meanings to be grounded and used?

Human communication is ostensive-inferential (Grice 1969, Wilson & Sperber 2004, Tomasello 2008, Sperber & Origgi 2010). In brief, this implies that the meaning of a given signal is not purely coded in the signal itself, but is inferred by the listener using both the signal and the context of communication as evidence (Sperber &

Origgi 2010). Sperber and Origgi provide an example sentence, 'It was too slow': an utterance which means little in isolation but, given enrichment by appropriate context and speaker-listener history, can hold a meaning as complex as 'the decrease in unemployment had been too slow in France when Jospin was Prime Minister to help him win the presidential election' or 'Jack's car was too slow (explaining why it was necessary to borrow Peter's)' (Sperber & Origgi 2010 p.124). Crucial to these enriched interpretations is the hearer's recognition of the speaker's intention to communicate: 'an intention to achieve a certain effect upon the mind of the hearer by means of the hearer's recognition of the very intention to achieve this effect' (Sperber & Origgi p.125).

The establishment of meaning via human communication, then, is not merely about matching signals to pre-existing cognitive concepts, or associating signals with frequently occurring situations. Rather, meaning inference crucially requires the hearer to be aware of the speaker's intention to communicate. This in turn makes the hearer peculiarly sensitive to any evidence the speaker might offer which gives clues to their intended meaning. This feature of ostensive-inferential communication has implications not only for the use of an established language, but for how the first human speakers would have grounded the first meanings (Scott-Phillips 2009). The dynamic interaction between what is salient to the hearer and the speaker, what prior concepts they might have, and what the hearer can infer from the pattern of evidence the speaker provides, creates a complex meaning space structured according to human communicative needs. Experiments that have shown the emergence of this kind of task-sensitive meaning space are Scott-Phillips's Embodied Communication Game, cited above, and Galantucci (2005)'s investigation of the emergence of graphical communication systems from a coordination task.

With these exceptions, the language evolution literature has not so far incorporated ostensive-inferential communication into investigations of meaning space emergence. In fact, most experiments and simulations have simply provided the meaning space as a publicly accessible and shared entity, rather than as a private space that may or may not fully correspond to the spaces of others. An exception to this latter tendency is Andrew Smith (2003), who examines how agents build up private meaning spaces and refine them through communication with other agents. One of Smith's major findings is that a 'clumpy world', where objects are grouped together in a systematic way, helps users' meanings to become more similar and

therefore allows a successful communication system to emerge and persist. Smith compares this 'clumpiness' to aspects of the physical world such as gravity. However, the 'clumpiness' that allows communication systems to get off the ground may not have to be fully inherent in the structure of the world, or in pre-linguistic concepts. It could partly be an artefact of the ways it makes sense to divide up the world for the purposes of communication.

Smith's work, however, does not model true inference: the setup of the agents' communication games means there is always a 'correct' real-world referent which is the intended meaning. The simulation therefore does not allow investigation of how the evidence offered for a speaker's intention can influence the meaning inferred. Xu & Tenenbaum (2007) show a concrete example of this effect: specifically, that word learners are sensitive to frequencies of evidence provided in the training input. The context with which a novel word was presented affected significantly whether both adult and child learners classified it as referring to a basic-level, subordinate or superordinate meaning. For example, the novel word 'fep', when presented with three pictures of Dalmatians, was interpreted as referring specifically to Dalmatians rather than to dogs in general, even though the evidence presented did not explicitly rule the latter hypothesis out. Xu & Tenenbaum call this 'the size principle', meaning that 'more specific meanings, with smaller extensions, are more likely than more general meanings, with larger extensions, when both are consistent with a given set of examples'. The authors describe this as a process of 'rational inductive inference' (Xu & Tenenbaum 2007a, abstract), taking into account two things: knowledge of the world, and the presumption that the teacher is being optimally helpful by providing a random sample from the word's potential referents (Xu & Tenenbaum 2007b, p.294).

I propose to use a method inspired by Xu & Tenenbaum to investigate experimentally how meanings are inferred and to what extent learners are sensitive to frequencies in the input. I aim to alter their methodology to establish whether the size principle affects inference when the training images presented are complex scenes. This shifts the locus of ambiguity from the superset/subset distinction to whether the intended meaning is the agent, the action, the patient, or some more complex combination of these elements. The aim will be to show that the process of inference itself defines individual meaning size, and that the meaning space as a whole is built up from how these episodes of inference compete and interlock over time. The resulting picture of the meaning space will look more akin to Ramscar et al. (2010)'s

vision of referents as flexible aspects of a probabilistic meaning inference process, rather than a pre-specified repertoire of cognitive concepts or a blind tracking of the recurrence of items across contexts.

This dissertation will first outline previous approaches to the human meaning space in the language evolution literature. Then, I will describe a plan for an experiment designed to test how human participants infer meanings across multiple contexts. I will then discuss the results in the light of the approaches to meaning inference sketched above, and conclude that words do not merely pair themselves with pre-existing cognitive concepts, or blindly associate with the items they recur with, but that the establishment of meaning is a dynamic process which is sensitive to the world knowledge of the hearer and their inference of the speaker's intention. I will then consider the implications of this result for the grounding of the first meanings and the evolution of language. The aim of the whole project is to deconstruct some assumptions about meaning space structure and to clear the ground for further investigation of the evolution of the human meaning space.

2. Background

2.1 Meaning spaces in the language evolution literature

As outlined in the Introduction, the language evolution literature has so far not paid a great deal of attention to how meaning spaces are established. More interest has focused on how words change over the process of iterated learning (Kirby & Hurford 2002, Kirby, Cornish & Smith 2008), or how graphical representations become symbolic via communicative interaction (Garrod et al. 2007, Theisen et al. 2010). Since what these researchers are primarily interested in is the cultural evolution of words or symbols, the meaning space in these experiments and simulations is planned in advance and kept constant.

The structure of this meaning space has varied. In Kirby & Hurford (2002), the meaning space was structured into predicates and arguments that could systematically recombine. In Kirby, Cornish & Smith (2008), the meaning space was structured along three dimensions (shape, colour and 'movement'), and was saturated, i.e. for each of the possible values of each parameter there was a corresponding referent. In Theisen et al. (2010), the referents to be identified shared salient semantic features in a structured way, along two dimensions of entity type and theme: for example, the referents 'teacher', 'school', and 'teaching' corresponded systematically to the referents 'firefighter', 'fire station' and 'fire-fighting'.

The results of the above simulations and experiments show that in conditions where the meaning space is fixed, shared and structured, the words or symbols that emerge are compositional: that is, elements of the signals correspond systematically to elements of the meanings. The exception in the above-cited work is Garrod et al. (2007). Here, the meaning space had some structure – referents were in themed groups, such as public buildings, actors, and household objects – but the structure was not completely systematic on the level of the other experiments. Specifically, it was not multi-dimensional or consistent (e.g. the group 'abstract' contained the referents 'loud', 'homesick' and 'poverty', which were not related in any other way than being abstract and did not correspond to elements of other groups). The resulting symbols exhibited limited compositional structure. These results imply that the emergence of compositionally structured words or symbols requires the pre-existence of a structured

meaning space. The question that has not yet been addressed in depth is whether such a meaning space corresponds accurately to the meaning space that underlies human language.

More recent work has started to question the assumption that the meaning space is uniformly structured. Work in progress by Mónica Tamariz and colleagues has introduced an idiosyncratic element, consisting of a unique 'appendage' for each item in a meaning space structured along the dimensions of shape and pattern (Tamariz et al., in prep.). This refinement effectively gives participants a choice on whether to attend to the structured elements of the meanings (presumably encouraging a compositional system) or to the idiosyncratic elements (presumably encouraging a non-compositional system, with an arbitrary label for each meaning). Preliminary results show that later generations of interacting participants tend to pay more attention to the structured elements of the meaning space. This suggests that structure in the meaning space could emerge dynamically from communication itself.

Even with variation in focus of attention, the meaning-signal pairs in the Tamariz study are still taught in isolation from context. This means that the experiment models a learning situation where both the meaning space itself, and the mappings between signals and meanings, are transparently accessible to the learner. A different experimental approach, focusing on word learning techniques rather than on the emergence of linguistic structure, places word learners in a situation of referential uncertainty, where exposures to each word are accompanied by multiple possible referents (K. Smith et al. 2011, Blythe et al. 2010). Learners are still capable of establishing meaning-signal pairings in this situation by using a technique called cross-situational learning (hereafter XSL). XSL learners track the potential referents that re-occur in contexts where a word is used, eliminating those that do not re-occur until only one candidate meaning is left. Human learners can use these kinds of strategies to learn realistically large lexicons (Blythe et al. 2010). Crucially, however, the combinations in which potential referents re-occur in these experiments are designed to allow complete disambiguation: i.e. there are no pairs of objects that always occur together, since this would make disambiguation between them impossible. The underlying assumption is that a number of discrete meanings exist in the meaning space, and that they will all at some point crop up without being accompanied by the same distractors.

While the XSL literature complicates the word-learning task by increasing the degree of referential uncertainty for the learner, it still assumes that the meaning-signal pair is publicly available. Andrew Smith has discussed the problems with transferring meanings directly along with signals, in that this makes the act of communication redundant (A. Smith 2003). A more realistic model is that each individual has their own private meaning space, mapping onto the objects in the world in an 'agent-specific internal semantic representation' (p.177).

As briefly outlined in the Introduction, Smith demonstrates that a successful communication system can emerge where agents map signals to their own private idiosyncratic meanings, developed through 'discrimination games' played on objects in the environment. Higher levels of meaning similarity between the agents lead to higher levels of communicative success (p.182). Levels of meaning similarity become much higher in a 'clumpy world', where groups of objects share identical values on one feature channel. Again, this result suggests that a meaning space that is structured in a particular way is crucial to communicative success. The model links the emergence of this structure explicitly to how the agents perceive the world, since the 'channels' are intended to represent sensory modalities.

Luc Steels and colleagues have undertaken similar work on the grounding of meanings, this time by embodied robotic agents viewing and communicating about real-world objects (Bleys et al. 2009, Steels et al. 2002). Their work has shown that agents' ontologies, or systems of meaning categories, can be aligned via successive communicative interactions grounded in the environment. The agents' conceptual categories change in 'a co-evolutionary process simultaneous with the construction of their lexical system' (Steels et al. 2002 p.253). The meanings the robots typically come up with are very different from the 'basic' meanings familiar with from human communication: 'for example, they may say "malewina" to mean [UPPER EXTREME-LEFT LOW-REDNESS]' (p.254). This example makes concrete Andrew Smith's assumption that the structure of the meaning space depends on the specific sensory modalities of the agent viewing it, and on what features it can count on being salient and discriminable for the hearer across a range of varying contexts. As in A. Smith (2003)'s simulation, the speaker's task is simply 'to draw attention through verbal means to an object (called the topic) in a visually-perceived reality' (Steels et al. 2002 p.260), and the hearer's task is to discriminate which object is intended from an array of several. The agents do this by tracking co-occurrences of a given word and its

meaning candidates, disambiguating when all but one candidate fails to appear in a communicative context. This is a kind of XSL, or more accurately cross-situational inference (since the meanings themselves are yet to be firmly established).

The results of Steels et al.'s grounded robotic experiments shed interesting light on the processes by which meanings are grounded and concepts formed, depending on what varies or stays the same across communicative contexts:

> Meanings that are compatible with the same situations will remain entangled until clear situations arise where they are different. This was for example the case for the word *bozopite*…There are two competing meanings: large area (large) and large width (wide). These meanings cooccur [sic.] often because objects that are large in area typically also have a large width. However when there are enough situations where the two are incompatible (for example because the object is very tall but not very wide), disambiguation starts and is enforced by the positive feedback loop. (Steels et al. 2002 pp.257-8)

Steels et al. describe the maintenance of these two meanings as polysemy, implying that there are objectively two distinct meanings in competition here. However, the structure of the experimental world partly determines the criteria for discrimination – e.g., there are some objects that are very tall but not very wide, allowing the meaning 'large width' to come into existence as an atomic concept. If all objects in the robots' world that had a large area also had a large width, and vice versa (i.e. large area and large width never, or very infrequently, occurred independently), there would be not be sufficient pressure for disambiguation. The word "bozopite" could continue meaning both 'large area and large width', with no requirement to separate these concepts or consider them as independent meanings. Another intriguing effect of the interaction between the nature of the robots' sensory apparatus and their communication system appears in the tendency for distinctions that are not easily shared to drop out of the language: 'very fine-grained distinctions based on subtle light variations would arise in the system but would have a hard time to propagate to the rest of the population because light conditions vary, even when the same set up is viewed from different angles' (Steels et al. 2002 p.258). The robots have identical sensory systems but differing perspectives, meaning their perceptions of the same scene can differ. A human analogue of the situation described above would be if a speaker were trying to draw a listener's attention to someone in a crowd, and described them using hair colour when the listener was behind a pillar and couldn't see the target's head.

The moral of the story is that if a distinction dividing a concept into two is not salient to both speaker and hearer – whether because the two things never occur separately, or because the distinction between them is not mutually manifest – two separate meanings will not be reinforced, because communication about this topic will not succeed. This principle could be extended beyond mutual ease of discrimination: it is possible that more generally, 'concepts which have no success in verbal interaction are not encouraged' (Steels et al. 2002 p.266).

Indeed, the lexicon is also determined by the definition of 'success in verbal interaction'. In the robot experiments, mutually salient distinctions only become lexicalised into meanings because the task assigned to the robots is to distinguish between the objects. Meanings are dependent not only on the structure of the world but also on the communicative needs of the interlocutors, as further illustrated by the Embodied Communication Game (Scott-Phillips et al. 2009). The game provides no pre-established meaning space or communication channel. However, the task of coordinating movements onto coloured squares means there is only a small range of meanings which can possibly be relevant. These meanings are flexible and sensitive to the current context of communication: for example, a meaning that starts along the lines of 'Not red!' can change to become 'Go to blue' once the players have agreed on a secondary default colour (p.229). The way meanings are established by successful pairs in this game is likely a better model of the way meanings are established in the real world: that is, by a probabilistic process highly sensitive to what has changed in the context since the last time the signal was used, and to the hearer's inference of the speaker's intention. Meanings in the real world are similarly not static, waiting for words to change around them, but are susceptible to change from the very nature of ostensive-inferential communication: 'in most communicative episodes, the speaker's original meaning and the hearer's reconstruction of the meaning will differ, at least to some extent' (A. Smith 2008 p.105). This paves the way for meaning change based on asymmetries between speaker and hearer understanding (Hoefler & A. Smith 2009).

2.2 Pre-existing concepts versus linguistically influenced concepts

Despite the evidence for the malleability of real-world meanings via linguistic processes, the argument that some or all concepts predate the acquisition of language has a long history. Whether in Fodor (1981)'s approach that all concepts are innate, or Harnad's 'hybrid' system where a connectionist network mediates the associations

between an innate structured symbol system and learned 'icons' (Harnad 1990), many theorists have come to the conclusion that some aspect of meaning space structure must be innate. In this view, the development of the lexicon is just a matter of matching words to pre-existing concepts. Paul Bloom makes this argument explicit: 'much of what goes on in word learning is establishing a correspondence between the symbols of a natural language and concepts that exist prior to, and independently of, the acquisition of that language' (Bloom 2000, p.242). Extending this idea from the learning situation of children to the pre-linguistic situation of a human ancestor, the assumption that objective, discrete concepts were 'out there' before the advent of language pervades the language evolution literature: for example in the noun- and verb-shaped 'pre-existing cognitive concepts' of Bickerton and Tallerman, cited in the Introduction.

It is intuitively appealing to imagine a pre-linguistic primate with concepts that correspond to the meanings we associate with common nouns and verbs. Jim Hurford makes a persuasive case in *The Origins of Meaning* (2007) for the possibility of animals forming 'pre-linguistic predicates' which '[correspond] to classes of input stimuli'. However, while this may work for noun-like meanings such as LEOPARD or visually identifiable verb-like meanings such as MOVE, there are some concepts corresponding to particular basic verbs that seem unlikely to be individuated cognitively by a non-communicative primate. To take, for example, the basic verb 'see': would a pre-linguistic primate necessarily unify their own visual sensory input with the idea of another's experience via the same organs under the same pre-linguistic concept? More likely seems the alternative: that it is useful to unite these experiences for the purposes of communication, and that this usefulness drove the establishment of the meaning for 'see'.

Of course, it could be claimed that the earliest atomic word meanings were those for whose conceptual existence in non-human primates there is already evidence: for example, Hurford (2007)'s LEOPARD, ROCK, MY-BABY CRY, etc. However, concepts which are already mutually cognitively individuated and salient are not necessarily those which would be most useful in early communication. A provocative possibility is that some meanings are, at least in part, created through communication, with the meaning taking the shape of the intersection between what the speaker wants to communicate and what the hearer is capable of inferring.

A theoretical framework for this idea is provided by Gentner and Boroditsky's hypothesis that 'relational systems', including verbs, are less cognitively transparent than simple nouns, and are therefore more likely to vary cross-linguistically (Gentner & Boroditsky, 2001). They argue that the language learner constructs her meaning space in the process of learning how a native language divides up these relational systems. Support for this hypothesis comes from research by Imai et al. (2008). In Chinese participants, children presented with a novel word mapped it to a noun referent in all conditions, whereas adults were more likely to map the word to a verb referent, both when the word was presented as a verb with arguments and when it was only presented as a bare word. Such differences in expectations regarding referents are established via learning a given native language. Therefore, linguistic experience in part forms the meaning space, rather than the entire meaning space existing a priori: 'inferring the denotations of verbs and other relational terms requires some knowledge of the language's semantic patterns' (Gentner & Boroditsky p.217). Gentner gives the example of a boy kicking a ball, where the resolution of a simple verb involves disentangling such varied relational associations as CONTACT (FOOT,BALL), SPIN(BALL), MOVE(BOY, FOOT), FLY(BALL). Every language will encode these associations somewhat differently. This is illustrated by the existence in some Mayan languages of so-called 'heavy verbs', which include specific arguments in their meaning: for example, in Tzeltal, 'the early verb *eat-tortilla* specifies both the event "eating" and the object "tortilla"' (Gentner & Boroditsky p.240).

The question, then, is what processes have led to these disjunctions in lexicalised concepts and therefore the shape of the meaning space across languages. From the literature summarised above, it seems likely that it has something to do with how the structure of events in the world interacts with the communicative intentions of the speakers. These processes will not always lead to meanings of some imagined prototypical noun or verb size, but may allow more complex meanings such as the Tzeltal heavy verb example. Andrew Smith (2008) argues that, given that meanings have to be reconstructed by inference, words with complex meanings are unlikely to be grounded, since 'the more complex and elaborate the semantic representation, the *less* likely the meaning can be faithfully reconstructed' (p.108). Jill Bowie (2008) elaborates that 'simpler elements of meaning…are more reliably inferred across different situations'. However, the size of the meaning reconstructed across contexts depends on the nature of these contexts: both in the sense of what candidate meaning

elements recur, and in the sense of the communicative purpose of the utterance. A full strategy for human meaning inference would therefore unite XSL-style occurrence tracking with world knowledge and intention-reading to make intelligent inferences about speaker meanings.

Xu & Tenenbaum's two 2007 papers, mentioned in the Introduction, display one approach to testing this united strategy in an experimental setting. They show that the inferences learners make about word meaning are not based only on blind tracking of frequencies or probabilities, but are sensitive to further structure in the input, in particular as it relates to world knowledge and assumptions about the teacher's helpfulness. The focus of the experiments is the distinction between subordinate, basic-level and superordinate levels of meaning represented by a given real-world object. For example, when presented with a word paired with a picture of a Dalmatian, despite the lack of overt competing meanings, the learner still has a choice between interpreting the meaning as subordinate (Dalmatian), basic (dog) or superordinate (mammal). Of course, there are further possibilities than these: a particular Dalmatian, pet, domesticated animal, animate being, as well as all the stranger meanings suggested by Quine (1960)'s 'gavagai' problem, for example, dog's ear, spotted pattern, loyalty, etc. However, Xu & Tenenbaum limit consideration to the subordinate/basic/superordinate levels by the way they construct the test set of images. The construction of the test set, and the task of picking out the referents that match the word that has been taught, are therefore themselves crucial to which meaning is inferred. As with Steels et al. (2002)'s robots, the criteria for discrimination and hence concept formation are based on what the target item has to be discriminated from. Xu & Tenenbaum design their test set to give participants a clear choice between subordinate, basic and superordinate levels of the target, with other distractors being from the unambiguously different domains of vegetables and vehicles.

The key manipulation Xu & Tenenbaum introduce is to vary the training set between 4 conditions. In condition 1, the learner is given a novel word paired with one image of a subordinate-level object (e.g. Dalmatian); in condition 2, three occurrences of a novel word paired with three images of a subordinate-level object (e.g. three Dalmatians); in condition 3, three occurrences of a novel word paired with one image of a subordinate-level object and two images of a basic-level object (e.g. a Dalmatian and two other dogs); in condition 4, three occurrences of a novel word

paired with one image of a subordinate-level object and two images of a superordinate-level object (e.g. a Dalmatian and two other mammals). The authors repeated this pattern across three domains of objects: Dalmatians/dogs/mammals, red peppers/peppers/vegetables, and yellow trucks/trucks/vehicles.

Their key finding was that learners are sensitive to 'suspicious coincidences' in the training set. For example, the novel word 'fep', when presented with one picture of a Dalmatian, was interpreted inconsistently as referring to Dalmatians or to dogs. However, when the novel word was presented with three pictures of Dalmatians, it was interpreted as referring specifically to Dalmatians rather than to dogs in general, even though the evidence presented did not explicitly rule the latter hypothesis out. Xu & Tenenbaum call this 'the size principle', meaning that 'more specific meanings, with smaller extensions, are more likely than more general meanings, with larger extensions, when both are consistent with a given set of examples' (Xu & Tenenbaum 2007b pp.291-2). The size principle depends on the learner's knowledge of the structure of the world the referent is drawn from. The learner assumes that the set presented with the word 'fep' is a representative sampling of the possible 'feps' in the world, and hence the conclusion drawn about the word's meaning is affected by the learner's knowledge of the world's structure as well as their ability to detect what Xu calls 'suspicious coincidences'. In the authors' words, 'pragmatic inferences based on communicative context affect generalizations about word meanings by changing the learner's probabilistic models' (Xu & Tenenbaum 2007a p.246). In their second paper, they establish that this effect only occurs in conditions where the learner is confident that the teacher has picked a deliberately helpful sample (Xu & Tenenbaum 2007b).

Xu & Tenenbaum's paradigm could be called 'cross-situational guessing'. In all training presentations, the level of meaning intended remains ambiguous, and the learner has to guess from the remaining possibilities (Dalmatian, dog, or mammal), working from the evidence they have been given. Their result shows that learners are sensitive enough to this evidence to be able to use it to go beyond the logical structure of the input. Xu & Tenenbaum's work also shows, in agreement with Callanan (1994), that the assumption that a basic-level meaning will be most successfully reconstructed across contexts of use is not necessarily valid. Indeed, while adults seemed to display a basic-level bias that cut across the results, children did not (Xu & Tenenbaum 2007a p.263). This suggests that 'a basic-level bias may not be part of the

foundations for word learning' (p.259), and may instead be, for example, an artefact of certain tendencies of child-directed speech.

2.3 Proposed experiment

I aim to use Xu & Tenenbaum's experimental paradigm to investigate further the ways in which learners use inference to determine word meanings. In particular, I want to extend the task to move beyond subordinate/basic/superordinate levels and into the arena of complex events. Rather than presenting learners with different combinations of subordinate, basic-level or superordinate-level referents, the experiment will present learners with images of events where participants stay the same or change across the training sets. This changes the arena of uncertainty on which 'cross-situational guessing' will have to work: learners will no longer have to decide between subordinate, basic and superordinate levels, but will have to determine whether the intended meaning refers to the action, the agent, the patient, or some combination of two or three. The results of the event-based experiment can then be compared directly with a replication of Xu and Tenenbaum's result in order to examine what might be different or similar about the way participants respond to events as opposed to objects alone.

Setting the experiment up as a word-learning task should encourage the assumption the size principle depends on, that the teacher is giving optimally helpful information (Xu & Tenenbaum 2007b). This assumption should encourage participants to process the information as optimally relevant. The size principle should then take effect as follows: where an event with the same agent, verb and patient is shown three times paired with the same novel word, a situation which would usually cause a learning failure in cross-situational learning (since there is no elimination of elements across learning episodes) could lead instead to inference of a meaning that includes all these elements. The motivation would be to model a real-world situation where events involving particular participants can recur with different frequencies, potentially affecting the size of particular meanings and hence overall meaning space structure.

Because Xu & Tenenbaum's methodology has not previously been extended to events, it is difficult to predict exactly what will happen. However, there are two main possibilities between which the results could fall. One possibility is that the size principle will apply equally to complex events as to subsets/supersets, and the

participants will be influenced by the 'suspicious coincidence' effect of being presented with, for example, three images of a shark eating a trout labelled with the same word to infer a meaning that includes all of these elements – effectively, learning that 'zinuw' means EAT(SHARK,TROUT). The interaction between the training sets presented and the learners' knowledge of the world should also apply here – i.e. if the learner knows that there are many eating events in the world that do not involve sharks and trout, and assumes that the training set they are presented with is a helpful random sample from the word's set of referents, the size principle should then encourage them to infer that the meaning of the word specifies all the elements the three images of the training set show. Another possibility is that the participants' expectations about word meanings, their tendencies in event interpretation, or even variations in their world knowledge, will counteract this effect, causing them to prefer to infer a noun or verb-sized meaning. Either way, the results will illuminate real-world inference and begin to unpack some assumptions about 'natural' meaning size.

3. Methods

3.1 Experiment 1. Replication of Xu & Tenenbaum (2007)

3.1.1 Participants

Participants were 21 postgraduate students (7 male, median age 23) from the University of Edinburgh, participating for a payment of £3. 12 were native English speakers and 9 were native speakers of other languages.

3.1.2 Materials

Disyllabic CVCVC non-words were generated using a random number generator, with any too similar to known English words eliminated.

Stimuli were digital colour photographs of 84 real-world objects. Images were sourced using Google Images, and were cropped and sized to 200x150 pixels.

The nature of the image set was changed slightly from the original Xu & Tenenbaum study. In their study, the three superordinate categories of mammals, vegetables and vehicles were reused once in each condition. With the study being set up as a 'word learning task', this could cause participants to exclude an already-assigned meaning when trying to determine the meaning of a novel word: in other words, mutual exclusivity effects will appear. Although participants could be instructed to treat each trial as unique and not to track information across trials, the tendency could still be to treat the experiment as a cross-situational learning task where they could make use of mutual exclusivity, e.g. 'I identified that word previously as referring to 'Dalmatian', so this one is more likely to mean 'dog''. Since the effects of interest were within trials, Xu and Tenenbaum's method was refined so that instead of re-using three superordinate categories, twelve unique superordinate categories were used, each of which only featured once in each trial. This difference makes Experiment 1 not strictly a replication, but this dissertation will continue referring to it as a replication for simplicity.

The set of images used in the first half of the experiment consisted of 12 sets of categories. On each trial, a novel word was presented with one or three images, in one of the four conditions detailed below:

Condition 1: One subordinate-level image

Condition 2: Three subordinate-level images

Condition 3: One subordinate-level image and two basic-level images

Condition 4: One subordinate-level image and two superordinate-level images

Three categories were assigned to each condition. The twelve categories, with their subordinate, basic-level and superordinate members, are presented in Table 1.

Table 1

Category	Subordinate	Basic	Superordinate	Condition
Fruit	Green apple	Red apple	Banana	1
		Yellow apple	Orange	
			Strawberry	
			Grapes	
Furniture	Deckchair	Armchair	Desk	1
		Kitchen chair	Wardrobe	
			Bookcase	
			Table	
Dairy products	Brie	Emmental	Milk	1
		Cheddar	Butter	
			Cream	
			Ice cream	
Mammals	Dalmatian	Alsatian	Cat	2
		Collie	Cow	
			Bear	
			Squirrel	
Vegetables	Red pepper	Green pepper	Leek	2
		Yellow pepper	Broccoli	
			Carrot	
			Tomato	
Vehicles	Blue mini	Grey Peugeot	Lorry	2
		Black smart car	Bicycle	
			Bus	
			Train	

Fish	Hammerhead shark	Whale shark Great White shark Tiger shark Bull shark	Clownfish Angelfish Anglerfish Stickleback	3
Office equipment	Green iMac G3	Four generic desktop computers	Printer Scanner Photocopier Shredder	3
Sports kit	Basketball	Cricket ball Baseball Football Rugby ball	Tennis racket Cricket bat Tennis net Football goal	3
Birds	Blue parrot	Red parrot Green parrot	Seagull Eagle Sparrow Robin Cormorant Puffin	4
Trees	Norway Spruce	Yew Pine	Copper beech Apple tree Oak Sycamore Cherry Willow	4
Drinks	Red wine	White wine Rosé wine	Whisky Gin Baileys Blue cocktail Pink cocktail Beer	4

Participants were first presented with a practice round of 3 trials to familiarise them with the experimental setup. Following this they were presented with 12 experimental trials as detailed above, always beginning with 3 examples of condition

1. This ordering followed Xu & Tenenbaum, and served to establish a 'baseline' for what participants would do when given only one piece of evidence rather than being able to rely on the greater clues provided by training sets of three. Conditions 2-4 then followed: participants were assigned to one of three possible orders that counterbalanced the sequence of the conditions.

For all trials, presentation of the training set was followed by a test array. This array always contained two subordinate-level matches, two basic-level matches, four superordinate-level matches, and sixteen distractors. Distractors were selected randomly from the full set of test images, with the exception that fruit stimuli did not appear when vegetables were the training set to avoid confusion (and similarly with mammals/birds).

Each trial presentation lasted 5 seconds; participants then had unlimited time to make their selections and continue to the next trial. Instructions for Experiment 1 were as follows:

> In this experiment, you will be taught a new language by being shown a word along with up to three examples of what the word means.
>
> You will see this training screen for 5 seconds. Then, you will be presented with an array of objects. Please click on those objects which you think match the meaning of the word you've just learnt.

3.2 Experiment 2. Extending Xu & Tenenbaum to complex events

The aim of Experiment 2 was to explore whether Xu & Tenenbaum's finding, that learners display an awareness of 'suspicious coincidence' that leads them to choose the most specific meaning extension consistent with the evidence given, would extend to learning words paired with complex events. Participants were as in Experiment 1.

Participants were presented with images of events involving an agent, an action and a patient, with the whole event paired with a word. The equivalent of Xu & Tenenbaum's condition 1 presented the learner with a novel word paired with one image of a specific agent doing a specific action to a specific patient (e.g. a koala climbing a ladder). Condition 2 presented three occurrences of a novel word paired with three images of a specific agent doing a specific action to a specific patient (e.g. three images of a shark eating a trout). Condition 3 presented three occurrences of a novel word paired with one image of a specific agent doing a specific action to a specific patient, and two images of different agents doing the same specific action to

the same specific patient (e.g. one image of Zorro riding a horse and two images of other people riding horses). Condition 4 presented three occurrences of a novel word paired with one image of a specific agent doing a specific action to a specific patient, and two images of different agents doing the same specific action to different patients (e.g. one image of a spider monkey hanging from a branch, and two images of other animals/people hanging from other objects).

The structure of the events is analogous to the subordinate, basic-level and superordinate structure of Xu and Tenenbaum's sample set in the following way. On the sense level, the concept of a Dalmatian includes within it the concept of a dog. However, on the reference level, the set of dogs includes within it the set of Dalmatians. Similarly, the event of a koala climbing a ladder, on the sense level, includes within it the concepts of climbing, the concept of a koala and the concept of a ladder. However, on the reference level, the set of climbing events contains within it the set of koala-climbing-ladder events. The more specific item with the smaller extension is thus the Dalmatian in the first case, and the koala climbing a ladder in the second case, since this involves all of the elements presented in training.

The structure and presentation of the events was analogous to Xu & Tenenbaum's study, as laid out in Table 2:

Table 2

	Agent	Action	Object
Subordinate	Same	Same	Same
Basic	Varies	Same	Same
Superordinate	Varies	Same	Varies

Therefore, the four conditions in the second part of the experiment were as follows:

Condition 1: One image of an agent performing an action on an object
Condition 2: Three images of the same agent performing the same action on the same object
Condition 3: One image of an agent performing an action on an object and two images of different agents performing the same action on the same object
Condition 4: One image of an agent performing an action on an object and two images of different agents performing the same action on different objects

The events presented were, as far as possible, conventional (i.e. the combinations of agent, action, and object were expected and familiar, rather than being counterintuitive). The possibility of using novel or counterintuitive events (although with familiar individual elements) was considered: for example, having a man throwing a hamster rather than a tennis ball, or a cat sitting on a miniature London bus. The idea was that the counterintuitive nature of the events would stop participants from automatically 'recognising' meanings that corresponded to words they were already familiar with. However, it was decided that using conventional events was preferable for two reasons. Firstly, it means participants should make sense of the event as an event, rather than simply looking for recurring elements in the images irrespective of their constituency. Secondly, an important factor in Xu & Tenenbaum's result is the world knowledge of participants: it is because they know that there are many other types of dog than a Dalmatian that the sample of three Dalmatians as opposed to three different dogs is suspicious. Therefore, if a participant is presented with three images of a particular person riding a horse, their awareness that many people ride, and that things other than horses can be ridden, is critical in their potential decision that the information they have been given leads towards a specific interpretation. If the participants were presented with completely novel events, their world knowledge would be violated by the presentation and so could not support the inferences they would go on to make.

Stimuli were designed to depict as clearly as possible an unambiguous agent, action, and patient for each event. Images were prepared using Google Images to locate suitable pictures of appropriate agents, actions and patients, splicing different images together where necessary. The agent, action, and patient were made as salient as possible by placing them on a white background. All images were presented at 200x150 pixels.

The test set for each trial was unique and consisted of the following 9 images:

- 2 subordinate matches – the same agent performing the same action on the same patient
- 2 basic-level matches – a different agent performing the same action on the same patient
- 2 superordinate-level matches – a different agent performing the same action on a different patient
- 1 image of the same agent performing the same action on a different patient

- 1 image of the same agent performing a different action on a different patient
- 1 image of a different agent performing a different action on the same patient

The latter three images were included to make room for a participant's hypothesis that the word referred to the agent + action, the agent alone, or the patient alone. The only potential meaning not represented in the test set was the hypothesis that the meaning of the word was a combination of agent + patient. This was excluded for reasons of practicality (it being difficult to find examples of agent + patient combinations which could be used both for the 12 actions in the training set and 12 additional novel actions) and because the hypothesis of an agent + patient meaning, while possible to draw from some of the training sets, was considered unlikely.

The number of superordinate-level matches (i.e. matches where the only consistent element was the action) was lower in this part of the experiment (2) than in the first part (4). This was done to avoid boosting any expectation of the participants that the word was a verb. The overall number of distractors was also lower: this was for reasons of space and clarity of presentation, since the complexity of the task was higher.

A complete list of the images presented, with conditions, is shown in Table 3.

In initial trials, a concern was raised that participants might be disposed to a noun bias given that the referents in the first part of the experiment were all nouns. To try and avoid this, participants were instructed specifically to learn the meaning of the word based on the evidence given. Instructions given to participants in Experiment 2 were as follows:

> This time, you will be shown a word along with one or three scenes.
>
> Your job is to learn the meaning of the word based on the evidence you are given, and then to select those scenes that you think the word applies to.

Table 3 (on next page)

Action	Subordinate	Basic	Superordinate	Agent+Action	Agent	Patient	Condition
Punching	Amir Khan punching a punchbag	A girl punching a punchbag	A female boxer punching The Joker	Amir Khan punching another boxer	Amir Khan holding a flag	A woman hugging a punchbag	1
		A woman punching a punchbag	A female boxer punching a chainsaw-wielding maniac				
Climbing	A koala climbing a ladder	A businessman climbing a ladder	A model climbing a rockface	A koala climbing a branch	A koala sleeping in a tree	A man carrying a ladder	1
		A fireman climbing a ladder	A woman climbing a cliff				
Chasing	A Border collie chasing a fox	A boy chasing a fox	A man chasing a woman	A Border collie chasing a rabbit	A Border collie biting a frisbee	A woman holding a fox	1
		A man chasing a fox	A boy chasing a man				
Kicking	David Beckham kicking a football	A girl kicking a football	A woman kicking a dog	David Beckham kicking a bottle	David Beckham holding a football shirt	A boy holding a football	2
		A woman kicking a football	A child kicking a tortoise				
Eating	A shark eating a trout	A man eating a trout	A boy eating a croissant	A shark eating a diver	A shark jumping out of water	A hand holding a trout	2
		A woman eating a trout	A woman eating broccoli				
Sitting	A cat sitting on a red cushion	A man sitting on a red cushion	A boy sitting on a stool	A cat sitting on a pavement	A cat catching a leaf	A girl hugging a red cushion	2
		A woman sitting on a red cushion	A girl sitting on an armchair				
Lifting	Chen Yanqing lifting weights	A man lifting weights	A man lifting a plank	Chen Yanqing lifting a dog	Chen Yanqing holding flowers	A man leaning on weights	3
		An orangutan lifting weights	A domo lifting a banana				
		A little girl lifting weights					
		An older man lifting weights					
Riding	Zorro riding a horse	Four different people riding horses	A woman riding a motorbike	Zorro riding a motorbike	Zorro brandishing a sword	A woman feeding a horse	3
			A man riding a bike				
Throwing	Andre Agassi throwing a tennis ball	Four different people throwing tennis balls	A man throwing a child	Andre Agassi throwing a coin	Andre Agassi kissing Steffi Graf	Rafael Nadal hitting a tennis ball	3
			A woman throwing a bouquet				
Jumping over	A kangaroo jumping over a gorse bush	A tiger jumping over a gorse bush	A dog jumping over a plant	A kangaroo jumping over a rock	A kangaroo carrying a joey	A bird perching on a gorse bush	4
		A man jumping over a gorse bush	A horse jumping over a bucket				
			A girl jumping over a gate				
			A squirrel jumping over a tuft				
Hanging	A spider monkey hanging from a branch	A girl hanging from a branch	A gibbon hanging from a net	A spider monkey hanging from a rope	A spider monkey hugging its baby	A red panda lying on a branch	4
		A bat hanging from a branch	A sloth hanging from a climbing frame				
			A frog hanging from a leaf				
			A man hanging from a ladder				
Diving	A kingfisher diving into a pond	A man diving into a pond	A woman diving into the sea	A kingfisher diving into an ice-hole	A kingfisher mating with another kingfisher	Ducks swimming in a pond	4
		A pelican diving into a pond	A penguin diving into a bucket				
			A seal diving into a waterfall				
			A man diving into a swimming pool				

4. Results

4.1 Data

Some data from participants 9 and 16 for Experiment 1 were excluded, since these participants did not initially understand that the task involved selecting more than one item. Participant 18's data for Experiment 2 were excluded because this participant frequently failed to choose the agent + agent + patient matches, appearing to show lack of understanding of the task. Participant 21's data were excluded because the participant did not realise at any point in the experiments that it was possible to select more than one item. Potential improvements in experiment design that could prevent these issues are suggested in the Discussion.

Following Xu & Tenenbaum, data were collapsed over the twelve superordinate categories and over the different test items within these categories. Participants chose distractors very rarely, less than 0.5% of the time. Therefore, distractor selection was not factored into any further analyses. No significant effect of Block (i.e. order in which participants saw the trials) was found (one-way ANOVA with Block as factor, all p-values > 0.05). Therefore Block was not factored into any further analyses.

4.2 Experiment 1

Participants were sensitive to the different patterns of evidence provided in each condition: in the one-example condition, participants selected all subordinate matches and, on average, around 50% of basic matches, whereas in the three-subordinate-example condition, participants selected only around 25% of the basic-level matches, $t(17) = 2.47$, $p < 0.05$. This result replicated the findings of Xu & Tenenbaum (2007) that 'suspicious coincidences' in the input make participants significantly less likely to generalise to basic-level referents.

The results of the replication are shown in Figure 1. It is important to bear in mind that selection of an example from any level implies that it is consistent to select all examples from the level below: for example, if a participant selects the basic-level examples of dogs, they ought also to select all the subordinate-level examples of Dalmatians unless they suffer a lapse in concentration. Hence, the subordinate-level examples are selected by all participants in all conditions, since they are consistent with all three levels of interpretation.

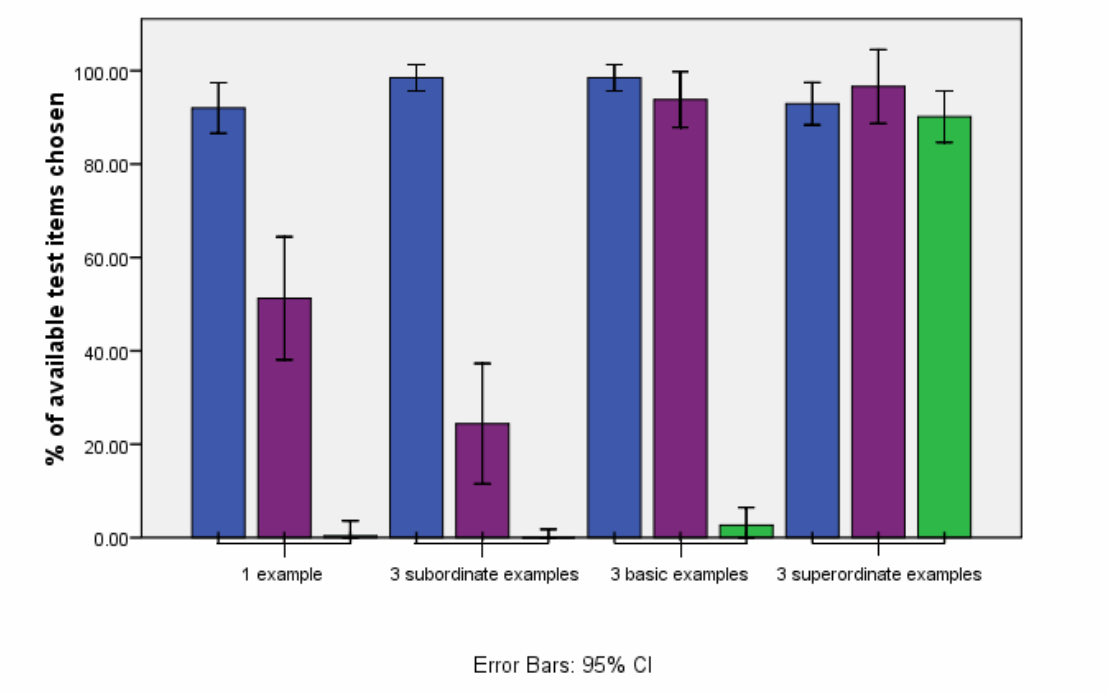

Figure 1

Results of Experiment 1. Blue: subordinate-level items (e.g. Dalmatian); purple: basic-level items (e.g. dog); green: superordinate-level items (e.g. mammal).

4.3 Experiment 2

Participants behaved differently in Experiment 2 (Figure 2). Unlike in Experiment 1, behaviour did not significantly differ between conditions 1 and 2: in particular, there was no difference in the percentage of 'basic' Action+Patient matches selected in the two conditions, $t(18) = 1.32$, $p = 0.2$. Participants generalised more to Action-consistent meanings in conditions 1, 2 and 3 than they generalised to superordinate meanings in the equivalent conditions in Experiment 1 (Condition 1: $T = 0$, $p < 0.005$, Condition 2: $T = 0$, $p < 0.005$, Condition 3: $T = 2$, $p < 0.05$). This suggests that a verb bias overlays the size principle. See Figure 3 for a summary of differences between Experiments 1 and 2.

As in Experiment 1, the selection of any given example implied the selection of any less complex meaning included within it: for example, if a participant selected Action-consistent examples, they ought also to select any examples containing the Action, i.e. Agent+Action+Patient, Agent+Action, and Action+Patient. Thus, as in Experiment 1, in all but a few peculiar instances to be raised in the Discussion, the Agent+Action+Patient examples are selected at 100% in every condition.

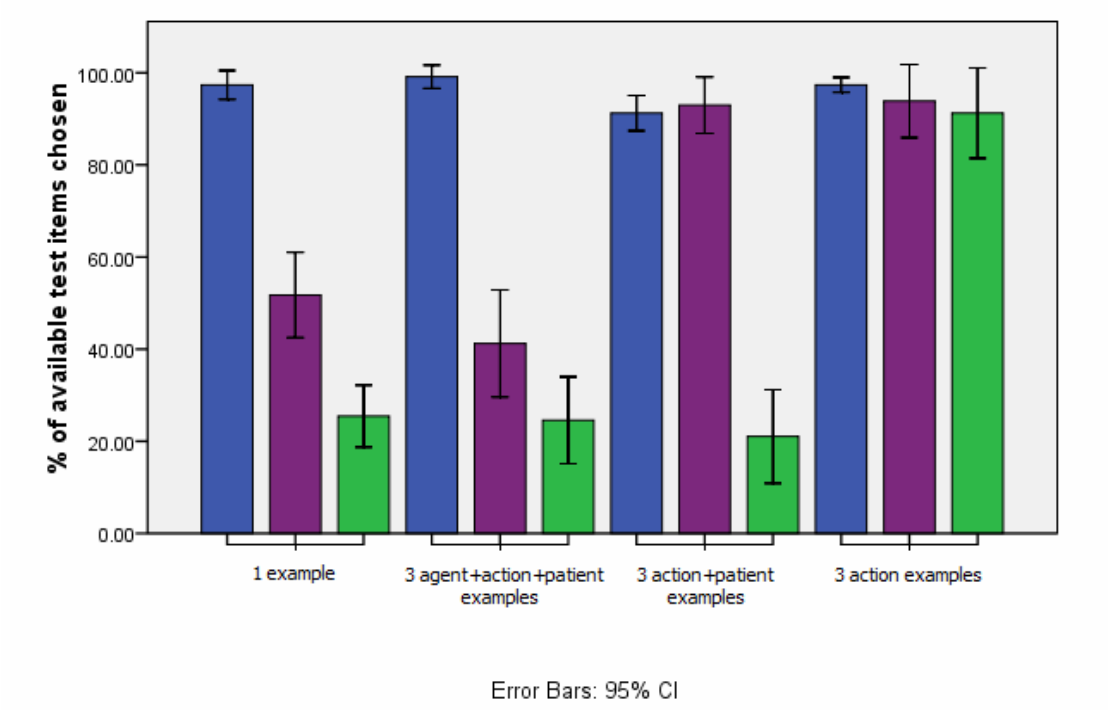

Figure 2

Results of Experiment 2. Blue: agent+action+patient-consistent items; purple: action+patient-consistent items; green: action-consistent items.

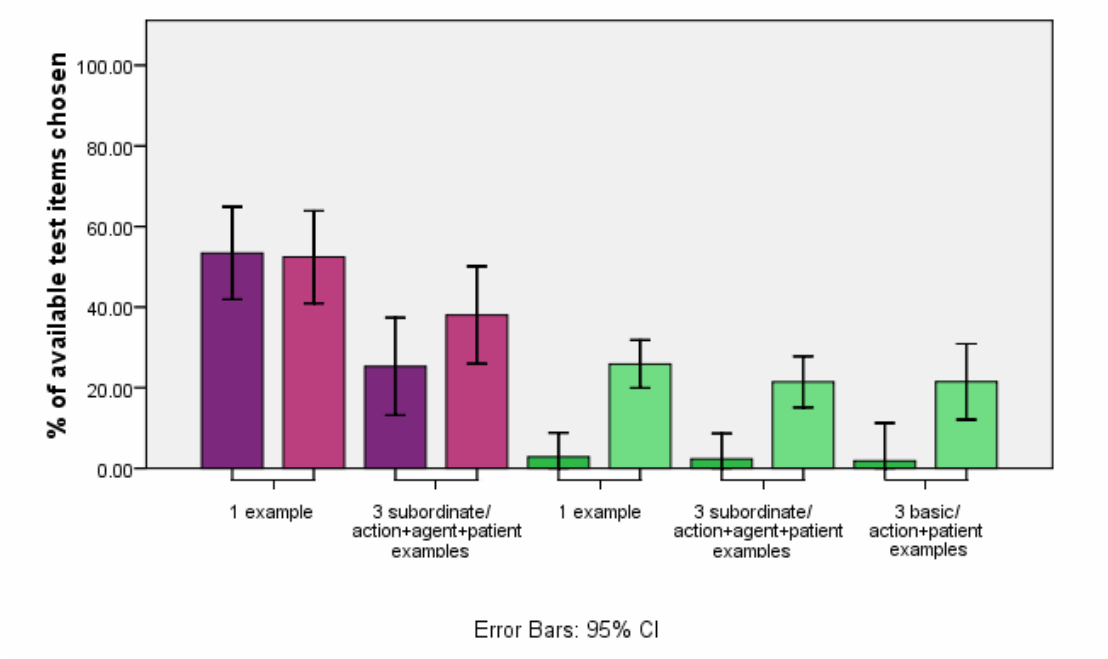

Figure 3

Summary of differences between Experiments 1 and 2. Purple: basic-level/action+patient-consistent items; green: superordinate-level/action-consistent items. Darker shade, Experiment 1, lighter shade, Experiment 2.

4.4 Effects of specific training and test items

Further examination of the data showed substantial variation in the effects produced by different trials. This suggests significant differences in how participants construed the structure of different object domains and events. The implications of the differences summarised below will be raised in the Discussion.

Basic-level matches chosen in the one-example condition of Experiment 1 varied significantly according to object domain, $\chi^2(2) = 19.5$, $p < .0001$. While there was no significant difference between the fruit and dairy trials, there was a significant difference between furniture and the other two trials: $Z = -3.217$, $p = .001$ (for furniture vs. dairy) and $Z = -2.810$, $p = .005$ (for furniture vs. fruit). Participants were generally less willing to select an armchair and a kitchen chair when presented with an image of a deckchair (16.7%, versus 75% and 61.1% for the other trials), suggesting that the structure of this domain in terms of subordinate, basic and superordinate levels differs significantly from the other domains. In condition 2, by contrast, there was no significant effect of trial ($\chi^2(2) = 5.33$, $p > .05$), suggesting that the additional weight of evidence from the size principle provided the participants with enough information to outweigh any natural variation in meaning divisions.

In Experiment 2, the uncertainty engendered by one-example training likewise created a significant effect of trial event on Action+Patient matches selected in condition 1, $\chi^2(2) = 14.44$, $p = .001$. Unlike in Experiment 1, however, there was also a significant effect of trial on the percentage of Action+Patient matches selected in condition 2, $\chi^2(2) = 10.429$, $p = .005$. Closer examination revealed that this effect was due to the percentage of Action+Patient matches selected for kicking (65.8%) being significantly higher than the percentage selected for either sitting (31.6%, $Z = -2.469$, $p = .014$) or eating (26.3%, $Z = -2.879$, $p = .004$). In fact, excluding the kicking scores, the difference between Action+Patient matches selected in conditions 1 and 2 is significant, similar to Experiment 1 ($Z = -1.969$, $p = .049$). It is possible, then, that the selection of trial Action here obscured a genuine size principle effect. The Discussion will further unpack the implications of these between-trial differences.

5. Discussion

5.1 Replication of Xu & Tenenbaum

The replication of Xu & Tenenbaum uncovered broadly the same results as the original experiment. Their headline result, that participants are significantly more likely to infer a subordinate-level meaning when presented with three subordinate-level examples than when presented with one example, held in the replication. However, the percentage of basic-level meanings inferred in the one-example condition was lower in the replication than in the original experiment. The percentage of basic-level matches chosen in the three-subordinate-example condition was also higher in the replication than the original experiment, making the overall difference, while still significant, smaller.

One potential explanation for the higher percentage of basic-level matches in condition 2 of the replication is that mutual exclusivity effects were operating in Xu & Tenenbaum's experiment, as hypothesised in the Methods. In both the original experiment and the replication, participants were presented first with three one-example trials, one from each object domain, and then with nine three-example trials counterbalanced by object domain and condition. However, in Xu & Tenenbaum's experiment, the three object domains of vegetables, mammals and vehicles were re-used throughout the experiment. The implication is that if a participant had decided on a basic-level meaning of 'dog' for the initial presentation of one picture of a Dalmatian, they would be less likely to infer the same meaning for a new word presented with three examples of Dalmatians, irrespective of the size principle. In the replication, a different object domain was used for each trial, eliminating the possibility of mutual exclusivity effects. This could in part account for the higher percentage of basic-level matches chosen in the three-subordinate-example condition, since participants would not have already assigned this level of meaning to a previously presented word in the one-example condition.

The lower percentage of basic-level matches chosen in the one-example condition, however, cannot be a consequence of this difference in methodology. In both experiments, the three one-example trials were presented first, eliminating the possibility of any mutual exclusivity effects at this stage. A likely alternative explanation is that the difference comes from characteristics of the object domains

used in each experiment. Xu & Tenenbaum use the same sets of mammals, vegetables and vehicles in all conditions, whereas the replication used three different domains for the one-example condition: fruit, dairy products and furniture. As reported in the Results, the furniture domain produced a significantly lower percentage of basic-level matches than the other two, pulling the overall percentage down. The exclusion of this category would bring the average percentage closer to Xu & Tenenbaum's. However, the point is not that furniture should not have been used as a domain, but that any set of results is going to be hugely dependent on the object domains and specific exemplars chosen for each meaning level. This point will be examined further in the discussion of inter-trial and inter-participant differences below.

5.2 Main differences between object and event results

Participants were more likely to be affected by the size principle when inferring meanings from object stimuli than from event stimuli. When presented with three examples of a complex event where agent, action and patient were the same, participants made overall similar meaning inferences as they did when presented with only one example. The pattern of inferences across subordinate and basic-level matches was similar to the pattern observed in the one-example condition in Experiment 1: participants picked all of the matches where agent, action and patient were consistent, and around 50% of the matches where action and patient were consistent.

Further differences appeared in the number of superordinate-level meanings inferred. In the event-based experiment, participants showed a higher tendency to infer a superordinate (i.e. action) meaning across all conditions. In the object-based experiment, inferences of superordinate-level meanings were at floor for all conditions where the training set was consistent with a more specific meaning. The size principle suggested by Xu & Tenenbaum would predict the same for events: that Action-only meanings would be inferred only in the condition where this was the most specific consistent meaning. Contrary to this expectation, participants inferred Action-only meanings around 25% of the time in conditions where a more specific meaning (Action+Patient or Agent+Action+Patient) was also consistent.

5.3 Objects versus events

What caused these differences between meaning inferences in object- and event-based trials? To answer this question, we need to unpack the differences between the object and event sets in the context of the experimental task.

For both the object- and event-based experiments, the structure of the training sets and test sets is designed to leave a certain level of uncertainty. This is what makes it a cross-situational guessing task, leaving room for inference. The kinds of uncertainty in the two experiments are different, however. In the object-based experiment, the construction of the test set defines the level of uncertainty as being between subordinate-level, basic-level and superordinate-level meanings. Even these categories are not absolute, but are constructed by the experimenters: for example, a differently-designed test set could offer 'dogs' as a subordinate-level meaning, 'domestic animals' as a basic-level meaning, and 'animals' as a superordinate-level meaning. Indeed, 'animal' may well be a more natural superordinate category for some participants than 'mammal'. The distractor images in this experiment are true distractors: if a participant selects one of them, it counts as an error showing either lack of understanding or a lapse in concentration.

For events, however, both the level of uncertainty defined by the test set and the nature of the distractors are different. The test set in Experiment 2 was designed to mirror Xu & Tenenbaum's test set, in that it had two matches for each level of meaning and then a number of distractors. However, depending on the training set of images, the inferences that can feasibly be made include a number of possibilities that include the distractor images, even if the size principle would potentially direct participants away from these meanings. For a novel word paired with either one or three examples of a koala climbing a ladder, for example, a participant could pick images from the test set consistent with any of the following possibilities: 'koala', 'climb', 'ladder', 'koala climb', 'climb ladder', or 'koala climb ladder'.

To allow this possibility, in addition to the Action, Action+Patient and Agent+Action+Patient matches cued by the training sets, the test array also contained images in which Agent+Action, Agent alone, or Patient alone was consistent. Only one test image was available for each of these possibilities, making the results vulnerable to skew when converted into percentages and making further analysis difficult. What is clear is that inference of Agent and Patient meanings was rare. Agent meanings were inferred in two trials in particular: Sit (where the cat was inferred as the Agent by 6 participants) and Climb (where the koala was inferred by 3

participants). For inference of Patient meanings, Lift followed by Kick contributed the most (4 participants and 2 participants respectively making the inference that the Patient was the intended meaning).

Why are Agent and Patient meanings inferred so rarely? One reason could be that the verb is the only thing consistent across all training sets, skewing against this conclusion. This possibility is discussed further below. However, an alternative contributing explanation has to do with the best way to exemplify actions versus agents or patients. To go back to Xu & Tenenbaum's assumption of the optimally helpful teacher: illustrating a transitive action requires including the agent and the patient to portray the action at all. By contrast, illustrating an agent or patient meaning is most optimally done by giving the agent or patient in isolation. This could be a factor preventing inference of agent or patient meanings, since too much excess information is provided to make this an optimally taught conclusion. This hypothesis could be tested by varying the training sets systematically between images of complex events that favour the interpretation of an agent meaning (i.e. where the only thing consistent is the agent) and images presenting the agent in isolation, to determine whether this affects the proportion of agent meanings inferred in each case.

Looking more closely at the two trials where the largest number of participants inferred an Agent or a Patient meaning ('Sit', where 6 participants inferred the cat/agent as the meaning, and 'Lift', where 4 participants inferred the weights/patient as the meaning): in these cases, the action included in the training images is intuitively closely associated with the agent or patient respectively. For a cat, sitting is a prototypical action; for a weight, being lifted is a prototypical action. This may work to make the action itself relatively 'invisible' to a learner, despite its presence across the training set.

The increased complexity of inference in the event-based experiment is not, however, only a matter of additional meaning possibilities. Participants' expectations about word meanings may also impact the event-based results more than the object-based results. In Xu & Tenenbaum's experiment, the meanings selected mostly corresponded to the referents of single English words (with the exception of 'red pepper' and 'yellow truck'). In the event-based experiment, only the agent, action, and patient meanings correspond to the referents of single English words. The fact that participants did not universally go for an action, agent, or patient meaning, even when presented with only one example to learn from, suggests that participants are

not just trying to match what they see with an existing referent of an English word. There is no English word that means, for example, PUNCH(MAN,PUNCHBAG) or even just PUNCH(PUNCHBAG). Despite this, PUNCH(PUNCHBAG) was specifically inferred as the meaning of a word paired with one image of a man punching a punchbag by 12 of 20 participants. However, it is possible that an expectation of a noun- or verb-like meaning for one word could be what holds participants back from inferring an Agent+Action+Patient meaning in the three-subordinate-example trial, creating the main difference between the object- and word-based experiments.

What the raw counts for the PUNCH(PUNCHBAG) example make clear is that the results are not clear-cut. Xu & Tenenbaum do not break down their results in this way, making it impossible to determine exactly where the 'gradedness' of generalisation comes from: that is, how participants in their one-example trials only pick 75% of basic-level matches and not 100% or 0%. The replication of Xu & Tenenbaum's result gave an opportunity to examine this in depth, as well as to investigate the source of the equivalent gradedness in events. What is clear is that the variation is not generally within-trial: when selecting from the test array, participants are usually consistent, and will either select all and only the Dalmatians, or all and only the dogs, rather than selecting one of two possible dogs. Further examination reveals two main sources of variation: between participants, and between trial objects or events.

5.4 The consequences of inter-participant differences

Close examination of the results of both experiments shows that participants vary widely in their classifications of objects and events. Indeed, the results suggest that participants may have differently structured meaning spaces depending on particular past experiences. Two examples of this effect in Experiment 1 are participants' responses to the trumpet/brass instrument/woodwind object domain (a 1-example practice trial) and the Brie/cheese/dairy product domain (a 1-example experimental trial). A participant who is a particular cheese connoisseur may be less likely to select all images of cheeses when given a word paired with one training image of Brie. Similarly, a participant who is a trumpet player is less likely to select a trombone and a sousaphone when shown a word paired with one example of a trumpet: they will specify that it means 'trumpet'. For the cheese example, 11 of 20 participants selected

all images of cheeses when shown one picture of Brie, leaving the remaining 9 purists who only selected the images of Brie.

The level of meaning inferred from one example image may be dependent not only on individual world knowledge, but also on how prototypical a given example is. In the construction of the training and test sets, care was taken to choose reasonably prototypical exemplars that would not cause confusion for participants: e.g. mammals picked for the test set were all terrestrial. It is possible that varying this would have an impact on the results, at least in the one-example trials. If given a novel word paired with one image of a 'prototypical' dog like a Dalmatian, are participants more likely to generalise to basic-level matches than if they were presented with one example of a less prototypical dog such as the Hungarian Puli (Figure 4)? Following the pattern of intelligent inference established by Xu & Tenenbaum, it would be interesting to carry out this experiment and see if a difference appeared. Xu & Tenenbaum's size principle relies on the assumption that a helpful teacher provides a random sample from the extension of the word. However, there could be an additional effect based not on pure frequency but on more complex cultural factors: it could be that the size principle includes the assumption that a helpful teacher would provide a prototypical dog if the basic-level meaning was intended, and that if the dog is unprototypical, the subordinate-level meaning must be intended instead.

Figure 4

Hungarian Puli (mid-flight).

This kind of inter-participant variation visibly affected selection of some test items in the event part of the experiment. Two participants, for example, did not select a particular test item of a woman climbing a wall, even though the rest of their selections implied that they had inferred the meaning 'climb'. The image was from a fashion shoot and the woman, unlike the one in the other 'climb' example, was not wearing professional climbing attire. This suggested that for two participants at least, appropriate clothing is part of the meaning of 'climb' when applied to people on rock-faces. Two participants also did not select a particular test item of a woman punching, even though their other selections implied an inference of 'punch', perhaps because the angle of the woman's arm was different from that of the exemplars given.
The caveat to taking too much from inter-participant differences, of course, is that we are discussing the behaviour of very few in each case, and that it is impossible to tell if the inconsistencies above are intentional on the part of the participants, or are just the result of lapses in concentration.

Some of the stimuli, however, yielded results that showed participants were clearly split on whether or not they were consistent with a given meaning. For example, in the 'throwing' test set, 5 of 20 participants selected an image showing Andre Agassi throwing a coin, while not selecting either the image of Andre Agassi kissing Steffi Graf or the images of other agents throwing other things, showing that they had not inferred an agent or action meaning. It seems that since the patient in the training images had been a tennis ball, these participants had inferred the meaning to be something like 'serving', and were willing to accept a serving-like arm posture even with a different object (a coin) as congruent with the meaning they'd inferred, rather than more general 'throwing' images of a bride with a bouquet or a man throwing a child into the air. This is also supported by the fact that 5 participants selected only one of the two test Action+Patient images of a person throwing a tennis ball: the favoured image featured an agent in a more prototypical 'serving' posture.

There were two further notable examples of stimuli that were interpreted in very different ways by different participants. The images intended to convey eating were particularly divisive, partly because this is an action difficult to convey in a static spliced image. 4 participants selected only one of the two Action+Patient matches, perhaps because of a combination of the non-intuitive nature of the pairing (person+trout) and the fact that the images that were sourced did not unambiguously

convey eating. A more interesting example came in participants' varied responses to the test images for 'riding'. 9 of 20 participants did not select one of the Agent+Action+Patient test items (featuring Zorro riding a horse). In this picture the horse was rearing up, seeming to show that for almost half the participants, a person sitting on a horse with a bipedal posture forms a recognisably different meaning than a person sitting on a horse with a quadrupedal posture. In fact, this difference is great enough to cause the percentage of Agent+Action+Patient meanings selected in condition 3 to be significantly lower than the equivalent in conditions 2 ($Z = -2.714$, $p = .007$) or 4 ($Z = -2.181$, $p = .029$).

The conclusion to be made from these different degrees of inter-participant variation is that the size principle does not act alone, and does not act consistently for every participant. Meaning spaces have a private, variable structure for each individual.

5.5 Differences between individual events

While the overall results show a reduced effect of the size principle for events in the lack of difference between conditions 1 and 2, further by-trial analysis complicates this picture. As shown in the Results, the first main difference found between the object experiment and the event experiment – that is, the higher number of basic-level matches chosen in the three-subordinate-example condition – rests on one trial verb, 'kick', which behaves significantly differently from the other two verbs in this condition, 'eat' and 'sit'. The verbs 'eat' and 'sit', when taken separately, appear to follow the pattern of Xu & Tenenbaum's original results, with participants showing sensitivity to the training input and inferring the most specific consistent meaning. That is, when shown three pictures of a shark eating a trout, participants tend to pick only pictures of a shark eating a trout from the test set, showing that they have inferred a word meaning along the lines of 'shark eats trout'. However, for 'kick', participants seem much less willing to infer an equivalently specific meaning, i.e. 'David Beckham kicks football'. A combination of factors could cause this. One could be the 'prototype' effect discussed in the context of dogs above: David Beckham is a prototypical footballer, and so this could encourage selection of a basic-level meaning (playing football generally) as opposed to a subordinate-level meaning (specifically David Beckham playing football). Another could be to do with the action-patient relationship: similarly to lifting weights, kicking is a prototypical action

for a football, and this could have contributed to making the specific Agent+Action+Patient combination less salient. For eating, the verb affected most strongly by the size principle, this effect could have come from a combination of looser connections between agent, action and patient (a lot of things eat other than sharks; sharks eat a lot of things other than trout), allowing participants to see clearly the 'suspicious coincidence' of the three elements staying constant across the training set. Just as participants in Xu & Tenenbaum's original study rely partly on their world knowledge of how Dalmatians relate to dogs, red peppers to peppers, etc., effects of particularly salient participant-action combinations may combine with the size principle to create varying results between different verbs in each trial.

Another possible contributing explanation for this effect, supported by divergent results for 'punch' as opposed to the other condition 1 verbs 'chase' and 'climb', could be a difference between human and animal agents. Perhaps participants are less willing to include a specific human agent in the inferred meaning of a given word than they are to infer a specific animal agent. However, these effects cannot be compared across conditions because of competing effects from the experimental manipulation. See below for an outline of how future work could investigate this.

5.6 The effect of word meaning expectations

Participants' native-language biases about the kind of things words typically refer to could also feasibly affect the meanings they are willing to infer. Participants may isolate one of the three elements of the scene in order to infer a noun or verb meaning, based on experience that tells them, for example, that there is no word that means 'ginger cat sits on a red cushion'. This could account for the 25% of action meanings inferred across the three conditions where the training set is consistent with a more specific meaning. This is supported by the fact that there was no effect of different trial verbs on the percentage of action meanings inferred in these conditions. Given the differing weights of evidence found in the three conditions, the lack of significant difference between them in terms of Action-consistent selections, and the lack of effects of specific trials, it is likely that this result reflects a systematic bias towards inferring a verb meaning.

It is difficult to establish for certain if a specific known bias is at work here, partly because most previous work has focused on, for example, whether learners are most likely to infer a noun meaning or a verb meaning (e.g. Imai 2008), rather than whether

learners would prefer to infer a simple noun- or verb-sized meaning over a meaning that incorporates all the elements of a complex event. Additionally, as discussed below, the training sets as a whole bias learners towards an action meaning, since the action alone stays constant across all sets no matter what the condition. However, participants still inferred a high proportion of complex meanings, involving at least two of agent, action and patient, across the various conditions. This suggests that word-meaning expectations did not completely override the effects of the size principle working on differently salient Agent+Action+Patient combinations.

5.7 Further avenues for investigation

There are several aspects of the current experiment which could be improved. As detailed at the beginning of the Results, three participants did not initially understand that the task involved selecting more than one item from the test array. The instructions could be improved to prevent this. Another refinement which would address both this problem and the issue of participant 18 not selecting subordinate-level matches, appearing to show lack of understanding, would be to institute a comprehension task with correct answers and unambiguous distractors before allowing participants to proceed to the experiment.

The method of measurement is also not ideal. Instead of using percentages of test items selected as the dependent variable, a more statistically satisfactory measurement would be raw counts, with the test set altered to ensure there were equal numbers of possibilities for each inferable meaning. This would also go towards undoing the bias towards particular meanings created by the structure of the test set, discussed further below.

For Experiment 2, in order to create a structure analogous to Xu & Tenenbaum's investigation into subordinate/basic/superordinate-level meanings, one of either agent, action or patient had to be chosen as the superordinate category. Likewise, one combination of two had to be chosen as the basic-level category. Action was chosen as superordinate and action + patient as basic-level, but this choice is arbitrary and weights participants' meaning expectations unfairly. In the current setup, the structure of the training sets taken as a whole, as well as the test set, effectively biases participants towards an action meaning. The action is the only thing that remains constant across training sets in all conditions; in addition, every test set contains 7

images consistent with an action meaning, as opposed to 4 consistent with an agent meaning and 5 consistent with a patient meaning.

An expanded version of the experiment, then, would ideally have 6 conditions. The designated basic and superordinate levels would vary across these conditions as shown in Table 4.

Table 4

	1	2	3	4	5	6
Subordinate	AVP	AVP	AVP	AVP	AVP	AVP
Basic	VP	AV	AV	AP	VP	AP
Superordinate	V	V	A	A	P	P

A= Agent, V = Action, P = Patient

This improved setup would allow further investigation of Gentner & Boroditsky (2001)'s hypothesis that actions are less cognitively transparent than objects, as well as the theory sketched above of how the assumption of a helpful teacher differently affects the teaching of actions versus agents/patients. If all 6 conditions were carried out, the differences between conditions where the action was superordinate and conditions where either the agent or patient was superordinate could well illuminate some of these potential effects. For example, would participants be more willing to infer an Action meaning than an Agent meaning, in conditions where the weight of evidence given for each was the same?

Another avenue that could be investigated is the effect of different expectations about word meanings on the results. Since the experiment is pitched as a word-learning task, participants' expectations about the typical referents of single words impinge on what meanings they are willing to infer: this, along with the overall bias towards action meanings noted above, seems to give rise to the 25% of action-consistent meanings inferred across the three conditions where the evidence allowed a more specific meaning. There are two potential ways to investigate to what extent expectations about typical word meanings impinge on the results. One would be to introduce a condition where, rather than one non-word being paired with the training examples, two or several non-words were used instead. Comparing these results with the one-word condition could illuminate how this impacts English speakers' inferences. An alternative avenue would be to explore differences between native

English speakers and native speakers of other languages. In particular, future work could systematically compare native speakers of languages that encode meanings lexically in a different way, for example, an agglutinative language such as Turkish, or a language with a greater tendency than English to form compounds, such as German.

The potential effects of human versus animal agents, noted above, could also be explored further by designing stimuli split between actions with human agents and actions with animal agents. The prototype effects discussed above could also be explored by varying the prototypicality of examples given to different participants.

The current experimental paradigm remains non-interactive, with only the illusion of interaction created by the "teacher"'s selection of training images influencing participants' inferences. Perhaps the most important avenue for future work would be to investigate how a real communicative situation impacts meaning inference. As discussed in the Introduction, the communicative task is hugely relevant both to individual meanings inferred and to the meaning space built up over successive interactions. A future experiment should try to model how the real-life 'task' of human communication dynamically creates its meaning space.

5.8 Conclusion

What we see in the complex pattern of the event-based results is consistent with the work of the size principle, overlaid by a combination of word meaning biases and action-, agent- and patient-specific effects. Some of these effects, in turn, vary hugely from participant to participant depending on their world knowledge. These results cast doubt on the idea that word learning is just about matching words to pre-existing cognitive concepts: it is highly doubtful that humans have a pre-existing concept equivalent to EAT(SHARK, TROUT), but the repeated co-occurrence of all these elements with a word, in a pattern at odds with the learner's world knowledge, can lead them to lexicalise this complex meaning.

More broadly, the results imply that there is unlikely to be such a thing as a uniform 'size' of verb or noun meaning that holds in all situations. As Mike Tomasello and colleagues argue in a study that uses novel nouns and verbs along with conventional actions, 'learning may even work differently for different types of nouns or verbs; for example, relational or complex nouns like *father* or *pedestrian* may behave more like verbs, and change-of-state verbs like *give* or *clean* might be easier

to learn than…characteristic actions' (Childers et al. 2002). Single words come into existence as single words because the meanings they express are useful in the interaction between the world and human communication, regardless of the complexity of these meanings. As Andrew Smith points out, a meaning like 'eat' can be seen as prototypically simple, but can alternatively be described by a detailed dictionary definition that incorporates many discrete concepts: 'to take into the mouth piecemeal, and masticate and swallow as food' (A. Smith 2008). However, the reason 'eat' exists as a basic verb, despite its complex implied set of agents (animate beings, perhaps even only those with mouths) and patients (edible substances), is because its referent is a highly salient human action, and all the implied elements that go along with it frequently co-occur in real communicative situations. The human meaning space, far from being a regular, stable structure, is variable and irregular, and the 'size' of a given meaning is determined by the preoccupations of the communicators who are using it. To refine Ramscar et al.'s point quoted at the beginning of this dissertation: referents are not discrete monolithic units which are static in size, but are more like variable points extending out from a humanly salient meaning centre.

This result casts a new light on the evolutionary questions about the grounding of the first meanings raised at the beginning of this dissertation. Evidence given for word meanings, interacting in a complex and nuanced way with world knowledge, works strongly enough in this experiment to override pre-established linguistic knowledge about the typical referents of single words. If this is the case even for participants who already have a fully-developed linguistic system, it implies that pragmatic factors determining meaning inference would have been even more influential without the scaffolding of a pre-established lexicon. The first meanings grounded were therefore not necessarily those already cognitively individuated by pre-linguistic primates. Neither were they simply evolved responses adapted to specific situations, along the lines of non-human primate calls: the picture of human meaning grounding that emerges from these results is radically different, involving rationally cross-referencing the elements that recur in the context of word use with world knowledge, and fundamentally based on recognising the speaker's intention to communicate. This implies that, as Tomasello (2008) and others have argued, the capacity to recognise and interpret intentions had to be present before language could evolve.

This capacity is not just necessary for an ostensive-inferential communication system to get off the ground, but also has a causative role in the complexity of

individual meanings grounded and therefore in the overall structure of the meaning space. In this experiment, the intentions of the 'teacher', as manifested in the structure of both the training sets and the test set, determine the meaning inferred in each trial. Similarly, in the Embodied Communication Game (Scott-Phillips et al. 2009), the participants end up with a meaning space structured around colour terms not simply because the colours are there to refer to, but because the successive demands of the task circumscribe these aspects of the environment as relevant, discriminable clues to partners' intentions. Ramscar et al. (2010)'s rethinking of reference, as a probabilistic and flexible aspect of predicting a speaker's intention, proves more helpful here than the traditional idea of a meaning space as a rigid set of connections between words and referents. The picture we arrive at via this approach looks very different from Stephen Pinker (1994)'s description of language as a system by which we can 'reliably cause precise new combinations of ideas to arise in each other's minds': these ideas and the way in which they combine are less precise and reliable, and much more sensitive to context and world knowledge, than they intuitively seem. This in turn cues a different way of thinking about the evolutionary processes that led to the current state of links between meanings and signals in human language. It could be that the meaning space as we know it emerged and attained its structure as a by-product of humans' attempts to read each other's intentions.

The picture these results create, of the dynamic establishment of meanings in a relevance-driven process sensitive to speakers' intentions, challenges assumptions about the pre-linguistic structure of the meaning space. Meanings of different complexity are more or less likely to be lexicalised depending on the interaction between the context of word use, the hearer's world knowledge, and the evidence provided for the speaker's intention. This reassessment of the idea of the meaning space has implications for the language evolution literature. If an ostensive-inferential process of meaning grounding makes the structure of the meaning space variable and at least partly idiosyncratic to different individuals, this complicates the much-addressed question of how linguistic structure arises. Recent work has shown that for highly structured meaning spaces, linguistic structure evolves to match meaning structure (Tamariz, in press). If there is no single objectively structured meaning space that exists before the advent of language, the next question to address is whether communication itself could structure the meaning space to a sufficient extent to enable compositional language to evolve. The natural direction for future work, as

discussed above, is to examine exactly how the task of human communication causes and permits variation in individuals' meaning spaces, while still allowing for the emergence of a shared and structured language.

6. References

Bickerton, D., 1990. *Language and Species*. Chicago: University of Chicago Press.

Bleys, J., Loetzsch, M., Spranger, M., Steels, L. 2009. The grounded colour naming game. In *Proceedings Spoken Dialogue and Human-Robot Interaction workshop at the RoMan 2009 conference*.

Bloom, P. 2000. *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Blythe, R.A., Smith, K. and Smith, A.D.M. 2010. Learning times for large lexicons through cross-situational learning. *Cognitive Science* 34(4), 620-642.

Burling, R. 2005. *The Talking Ape: How Language Evolved*. Oxford: Oxford University Press.

Childers, J.B., Tomasello, M., 2002. Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Developmental Psychology* 38(6), 967-978.

Fodor, J.A. 1981. The present status of the innateness controversy. In Jerry Fodor (ed.), *Representations*. Cambridge, MA: MIT Press, 283-92.

Galantucci, B., 2005. An experimental study of the emergence of human communication systems. *Cognitive Science* 29, 737–767.

Garrod, S., Fay, N., Lee, J., Oberlander, J., MacLeod, T., 2007. Foundations of representation: where might graphical symbol systems come from? *Cognitive Science* 31(6), 961-987.

Gentner, D., & Boroditsky, L., 2001. Individuation, relativity and early word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development*, pp. 215–256. Cambridge, UK: Cambridge University Press.

Golinkoff, R., Chung, H., Hirsh-Pasek, K., Liu, J., Bertenthal, B., & Brand, R., 2002. Young children can extend motion verbs to point-light displays. *Developmental Psychology*, 38, 604–615.

Grice, P. 1969. Utterer's meaning and intention. *The Philosophical Review* 78 (2), 147-177.

Harnad, S. 1990. The symbol grounding problem. *Physica D* 42, 335-346.

Hauser, M. D., Chomsky, N., Fitch, W.T. 2002. The faculty of language: what is it, who has it, and how did it evolve?. *Science* 298(5598), 1569-1579.

Hoefler, S.H., Smith, A.D.M., 2009. The pre-linguistic basis of grammaticalisation: a unified approach to metaphor and reanalysis. *Studies in Language* 33(4), 883-906.

Hurford, J. 2007. *The Origins of Meaning: Language in the Light of Evolution*. Oxford: Oxford University Press.

Imai, M., Li, L., Haryu, E., Okada, H., Hirsh-Pasek, K., Golinkoff, R.M., Shigematsu, J., 2008. Novel Noun and Verb Learning in Chinese-, English-, and Japanese-Speaking Children. *Child Development* 79(4), 979–1000.

Jackendoff, R. 1999. Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences* 3(7), 272-279.

Keil, F.C., 1989. *Concepts, kinds, and cognitive development*. MIT Press, Cambridge, MA.

Pinker, S. 1994. *The Language Instinct*. New York: HarperCollins.

Quine, W. V. O. 1960. *Word and object*. Cambridge, MA: MIT Press.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., Thorpe, K. 2010. The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science* 34, 909–957.

Scott-Phillips, T. C., Kirby, S., Ritchie, G. R. S. 2009. Signalling signalhood and the emergence of communication. *Cognition* 113(2), 226-233.

Scott-Phillips, T. C. 2010. The evolution of communication: Humans may be exceptional. *Interaction Studies* 11(1), 78-99.

Seyfarth, R.M., Cheney, D.L., Marler, P. 1980. Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science* 210(4471), 801-803.

Smith, A.D.M. 2003. Intelligent meaning creation in a clumpy world helps communication. *Artificial Life* 9(2), 175-190.

Smith, A.D.M., 2005. The inferential transmission of language. *Adaptive Behavior* 13(4), 311-324.

Smith, A.D.M. 2008. Protolanguage reconstructed. *Interaction Studies* 9(1), 100-116.

Smith, K., Smith, A.D.M., Blythe, R.A. 2011. Cross-situational learning: an experimental study of word-learning mechanisms. *Cognitive Science* 35(3), 480-498.

Sperber, D., Origgi, G. 2010. A pragmatic perspective on the evolution of language. In R. K. Larson, V. Déprez & H. Yamakido (eds.) *The Evolution of Human Language: Biolinguistic Perspectives*, Cambridge: Cambridge University Press, 124-131.

Steels, L. 1999. How language bootstraps cognition. In Wachsmutt, I. and Jung, B., editor, *KogWis '99: Proceedings der 4. Fachtagung der Gesellschaft für Kognitionswissenschaft*, Braunschweig: Infix, 1-3.

Steels, L., Kaplan, F., McIntyre, A., & Van Looveren, J. 2002. Crucial factors in the origins of word-meaning. In A. Wray (Ed.), *The transition to language*. Oxford: Oxford University Press, 252-271.

Tallerman M. 2007. Did our ancestors speak a holistic protolanguage?. *Lingua* 117(3), 579-604.

Tamariz, M. (in press). Linguistic structure evolves to match meaning structure. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp.). Austin, TX: Cognitive Science Society.

Tamariz, M., Roberts, S., Cornish, H., Smith, K., Kirby, S. In preparation.

Theisen, C., Oberlander, J. and Kirby, S. 2010. Systematicity and arbitrariness in novel communication systems. *Interaction Studies* 11(1),14-32.

Tomasello, M. 2008. *Origins of Human Communication*. Cambridge, MA: MIT Press.

Wilson, D., Sperber, D. 2004. Relevance theory. In Horn, L.R. & Ward, G. (Eds.) *The Handbook of Pragmatics.* Oxford: Blackwell, 607-632.

Wray, A., 2000. Holistic utterances in protolanguage: the link from primates to humans. In Knight, C., et al. (Eds.). *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form.* Cambridge: Cambridge University Press, 285–302.

Wray, A., 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Xu, F., Tenenbaum, J.B. 2007a. Word learning as Bayesian inference. *Psychological Review* 114(2), 245–272.

Xu, F., Tenenbaum, J.B. 2007b. Sensitivity to sampling in Bayesian word learning. *Developmental Science* 10:3, 288–297.